J. Mol. Biol. (1966) **19**, 548–555

Codon–Anticodon Pairing:

The Wobble Hypothesis

F. H. C. Crick

Medical Research Council, Laboratory of Molecular Biology

Hills Road, Cambridge, England

(*Received 14 February 1966*)

It is suggested that while the standard base pairs may be used rather strictly in the first two positions of the triplet, there may be some wobble in the pairing of the third base. This hypothesis is explored systematically, and it is shown that such a wobble could explain the general nature of the degeneracy of the genetic code.

Now that most of the genetic code is known and the base-sequences of sRNA molecules are coming out, it seems a proper time to consider the possible base-pairing between codons on mRNA and the presumed anticodons on the sRNA.

The obvious assumption to adopt is that sRNA molecules will have certain common features, and that the ribosome will ensure that all sRNA molecules are presented to the mRNA in the same way. In short, that the pairing between one codon–anticodon matching pair will to a first approximation be "equivalent" to that between any other matching pair.

As far as I know, if this condition has to be obeyed, and if all four bases must be distinguished in any one position in the codon, then the pairing in this position is *highly likely* to be the standard one; that is:†

$$\begin{aligned} & \text{G} = = = = \text{C} \\ \text{and} \quad & \text{A} = = = = \text{U} \end{aligned}$$

or some equivalent ones such as, for example,

$$\begin{aligned} & \text{I} = = = = \text{C} \\ \text{and} \quad & \text{A} = = = = \text{T} \end{aligned}$$

since this is the only type of pairing which allows all four bases to be distinguished in a strictly equivalent way.

We now know enough of the genetic code to say that in the *first two* positions of the codon the four bases are clearly distinguished; certainly in many cases, and probably in all of them. I thus deduce that the pairings in the first two positions are likely to be the standard ones.

† Throughout this paper the sign $= = = =$ is used to mean "pairs with". If two bases are equivalent in their coding properties, this is written $\begin{matrix}\text{U}\\\text{C}\end{matrix}$ or $\left.\begin{matrix}\text{U}\\\text{C}\end{matrix}\right\}$

However, what we know about the code has already suggested two generalizations about the third place of the codon. These are:

(1) U, C † this already appears true in about a dozen cases out of the possible 16, and there are no data to suggest any exceptions.

(2) A, G probably true in about half of the possible 16 cases, but the evidence suggests it may perhaps be incorrect in several other cases.

The detailed experimental evidence is rather complicated and will not be discussed here. (For details of the code see, for example, Nirenberg *et al.*, 1965; and Söll *et al.*, 1965.) It suffices that these rules *may* be true, as suggested by Eck (1963) a little time ago. Alternatively, only the first one may be true.

This naturally raises the question: Does *one* sRNA molecule recognize more than one codon, e.g. both UUU *and* UUC. Some evidence for this was first presented by Bernfield & Nirenberg (1965). They showed that *all* the sRNA for phenylalanine can be bound by poly U, although this sRNA also recognizes the triplet UUC, at least in part. More recent evidence along these lines is presented in Söll *et al.* (1966) and Kellogg *et al.* (1966). Again I do not wish to discuss here the evidence in detail, but simply to ask: If one sRNA codes both XYU and XYC, how is this done?

Now if we do not know anything about the geometry of the situation, it might be thought that almost any base pairs might be used, since it is well known that the bases can be paired (i.e. form at least two hydrogen bonds) in many different ways. However, it occurred to me that if the first two bases in the codon paired in the standard way, the pairing in the third position might be *close* to the standard ones.

We therefore ask: How many base pairs are there in which the glycosidic bonds occur in a position close to the standard one? Possible pairs are:

$$\text{G} ==== \text{A} \tag{1}$$

In my opinion this will not occur, because the NH_2 group of guanine cannot make one of its hydrogen bonds, even to water (see Fig. 1).

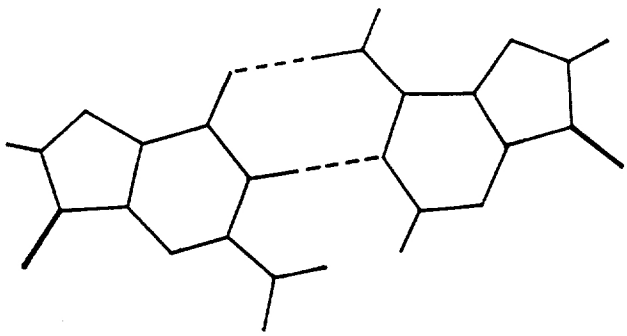

FIG. 1. The unlikely pair guanine–adenine.

$$\text{U} ==== \text{C} \tag{2}$$

This brings the two keto groups rather close together and also the two glycosidic bonds, but it may be possible (see Fig. 2).

† This symbol implies that both U and C code the same amino acid.

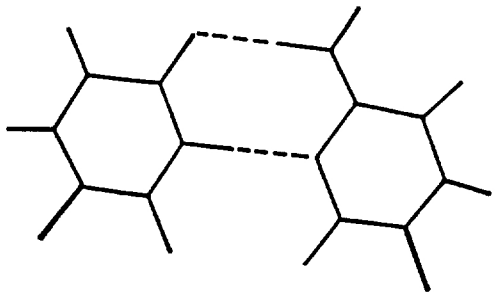

FIG. 2. The close pair uracil–cytosine.

$$U ==== U \tag{3}$$

Again rather close together (see Fig. 3).

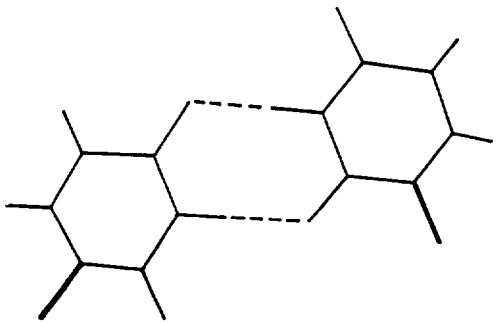

FIG. 3. The close pair uracil–uracil.

$$G ==== U \tag{4}$$

$$\text{or } I ==== U$$

These only require the bond to move about 2·5 Å from the standard position (see Fig. 4).

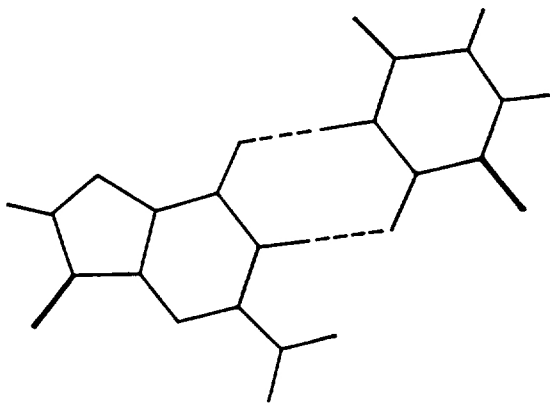

FIG. 4. The pair guanine–uracil (the pair inosine–uracil is similar).

$$I ==== A \quad (5)$$

This is perfectly possible. Poly I and poly A will form a double helix. The distance between the glycosidic bonds is increased (see Fig. 5).

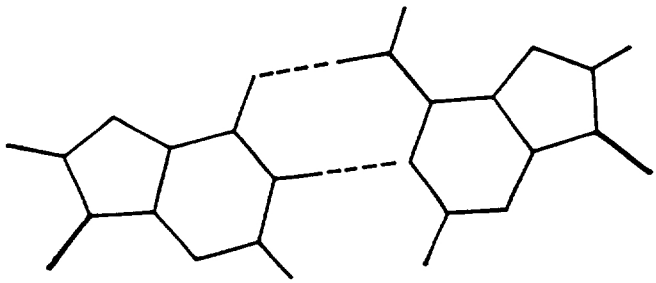

FIG. 5. The pair inosine–adenine.

As far as I know, these are all the possible solutions if it is assumed that the bases are in their usual tautomeric forms.

I now postulate that in the base-pairing of the third base of the codon there is a certain amount of play, or wobble, such that more than one position of pairing is possible.

As can be seen from Fig. 6, there are seven possible positions which might be reached by wobbling. However, it by no means follows that all seven are accessible, since the molecular structure is very likely to impose limits to the wobble. We should therefore strictly consider all possible *combinations of allowed positions*. There are 127 of these, but most of them are trivial. If we adopt the rule that *all four bases* on the codon (in the third position) must be recognized (that is, paired with) we are left with 51 different combinations. This is too many for easy consideration, but fortunately we can eliminate most of them by only accepting combinations which do not violate the broad features of the code. If we assume:

(a) that all four bases must be recognizable;

(b) that the code must *in some cases* distinguish between

$\left.\begin{matrix}U\\C\end{matrix}\right\}$ and $\left.\begin{matrix}A\\G\end{matrix}\right\}$ as it appears to do for the pairs

Phe	Tyr	His	Asn	Asp
Leu	C.T.†	Gln	Lys	Glu

(not all of which are likely to be wrong)

then by strictly logical argument it can be shown both that the standard position must be used, and that the three positions on the left of Fig. 6 cannot be used.

This leaves us with only four possible sites to consider one of which—the standard one—must be included. There are therefore only seven possible combinations. I have examined all these, but I shall restrict myself here to the case in which all four positions are used, as this is structurally the most likely and also seems to give the code (called code 4 in the note privately circulated) which best fits the experimental data.

† C.T., Chain termination.

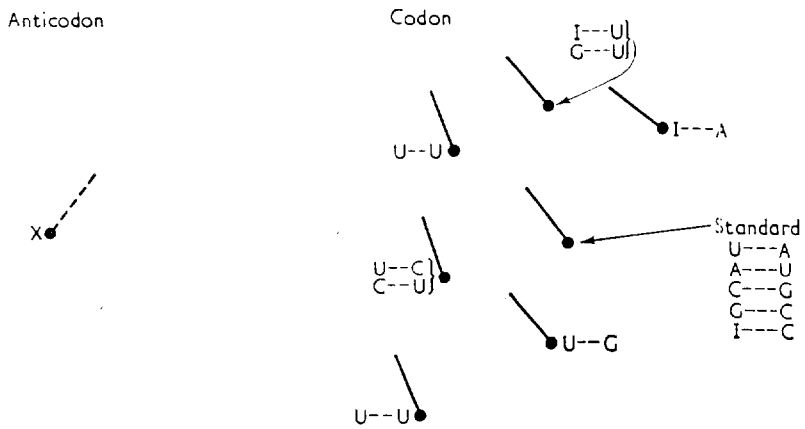

FIG. 6. The point X represents the position of the C_1' atom of the glycosidic bond (shown dotted) in the anticodon. The other points show where the C_1' atom and the glycosidic bond fall for the various base pairs. (Pairs with inosine in the codon have been omitted for simplicity.) The wobble code suggested uses the four positions to the right of the diagram, but not the three close positions.

The rules for pairing between the third base on the codon and the corresponding base on the anticodon are set out in Table 1. It can be seen that these rules make several strong predictions:

(1) it is not possible to code for either C alone, or for A alone.

For example, at the moment the codon UGA has not been decisively allocated. Wobble theory states that UGA might either:

(a) code for cysteine, which has UGU and UGC; or

(b) code for trypotophan, which has UGG; or

(c) not be recognized.

TABLE 1

Pairing at the third position of the codon

Base on the anticodon	Bases recognized on the codon
U	A, G
C	G
A†	U
G	U, C
I	U, C, A

† It seems likely that inosine will be formed enzymically from an adenine in the nascent sRNA. This may mean that A in this position will be rare or absent, depending upon the exact specificity of the enzyme(s) involved.

However it does *not* permit UGA to code for any amino acid other than cysteine or tryptophan. This rule could also explain why no suppressor has yet been found which suppresses only *ochre* mutants (UAA), although suppressors exist which suppress both *ochre* and *amber* mutants (UA^{A}_{G}).

(2) If an sRNA has inosine in the place at the relevant position on the anticodon (i.e. enabling it to pair with the third base of the codon), then it must recognize U, C and A in the third place of the codon. Conversely, those amino acids coded only by XY^{U}_{C} (such as Phe, Tyr, His, etc.) cannot have inosine in that place on their sRNA.

(3) Wobble theory does not state exactly how many different types of sRNA will actually be found for any amino acid. However if an amino acid is coded for by all four bases in the third position (as are Pro, Thr, Val, etc.), then wobble theory predicts that there will be at least two sRNA's. These can have the recognition pattern:

$$\left.\begin{matrix}U\\C\end{matrix}\right\} \text{ plus } \left.\begin{matrix}A\\G\end{matrix}\right\}$$

or

$$\left.\begin{matrix}U\\C\\A\end{matrix}\right\} \text{ plus } G$$

Note that the sets actually used for any amino acid may well vary from species to species.

The Anticodons

At this point it is useful to examine the experimental evidence for the anticodon. In the sRNA for alanine from yeast, Holley *et al.* (1965) have the following sequences:

——— pUpUpIp Gp CpMeIpΨp ———

position ——— 36 37 38 ———

Zachau and his colleagues (Dütting, Karan, Melchers & Zachau, 1965) have for one of the serine sRNA's from yeast:

——— pΨpUpIpGpApA^{+}pΨp ———

(A^{+} stands for a modified A)

For the valine sRNA from yeast, Ingram & Sjöquist (1963) have shown that the only inosine occurs in the sequence:

——— pIpApCp ———

Holley *et al.* (1965) have already pointed out that IGC is a possible anticodon for alanine, and the additional evidence makes it almost certain to my mind that this is correct, and that the anticodons are as given in the Table below†:

† *Note added 26 April 1966.* Drs J. T. Madison, G. A. Everett and H. Kung (personal communication) have completed the sequence of the tyrosine sRNA from yeast. The sequence strongly suggests that the anticodon in this case is GΨA, corresponding to the known codons UA^{U}. Since Ψ can form the same base pairs as U, this is in excellent agreement with the previous data.

Yeast sRNA

	Anticodon	Codon
Ala	I G C	G C ?
Ser	I G A	U C ?
Val	I A C	G U ?

remembering that the pairing proposed between codon and anticodon is *anti*-parallel. Thus I confidently predict: the anticodon is a triplet at (or very near) positions 36–37–38 on every sRNA, and that the *first two bases* in the codon pair with this (in an anti-parallel manner) *using the standard base pairs*.

However, inosine does not occur in every sRNA. In particular Holley *et al.* (1963) (and personal communication) have reported that the tyrosine sRNA has two peaks, neither of which contains inosine. Moreover, Sanger (personal communication) tells me that there is rather little inosine in the total sRNA from *E. coli*.

Testing the Theory

Two obvious tests present themselves:

(1) To find which triplets are bound by any one type of sRNA. This is being done by Khorana and his colleagues (Söll *et al.*, 1966), and also by Nirenberg's group (Kellogg, Doctor, Loebel & Nirenberg, 1966). The difficulty here is to be sure that the sRNA used is pure, and not a mixture.

(2) To discover unambiguously the position of the anticodon on sRNA, and to find further anticodons. This will certainly happen as our knowledge of the base sequence of sRNA molecules develops. The absence of inosine from any anticodon is obviously of special interest.

In conclusion it seems to me that the preliminary evidence seems rather favourable to the theory. I shall not be surprised if it proves correct.

I thank my colleagues for many useful discussions and the following for sending me material in advance of publication: Dr M. W. Nirenberg, Dr H. G. Khorana, Dr G. Streisinger, Dr W. Holley, Dr J. Fresco, Dr H. G. Zachau, Dr C. Yanofsky, Dr H. G. Wittmann, Dr H. Lehmann and Dr J. D. Watson.

REFERENCES

Bernfield, M. R. & Nirenberg, M. W. (1965). *Science*, **147**, 479.
Dütting, D., Karan, W., Melchers, F. & Zachau, H. G. (1965). *Biochim. biophys. Acta*, **108**, 194.
Eck, R. V. (1963). *Science*, **140**, 477.
Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R. & Zamir, A. (1965). *Science*, **147**, 1462.
Holley, R. W., Apgar, J. Everett, G. A., Madison, J. T., Merrill, S. H. & Zamir, A. (1963). *Cold Spr. Harb. Symp. Quant. Biol.* **28**, 117.
Ingram, V. M. & Sjöquist, J. A. (1963). *Cold. Spr. Harb. Symp. Quant. Biol.* **28**, 133.

Kellogg, D. A., Doctor, B. P., Loebel, J. E. & Nirenberg, M. W. (1966). *Proc. Nat. Acad. Sci., Wash.* **55**, 912.
Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F. & O'Neal, C. (1965). *Proc. Nat. Acad. Sci., Wash.* **53**, 1161.
Söll, D., Jones, D. S., Ohtsuka, E., Faulkner, R. D., Lohrmann, R., Hayatsu, H., Khorana, H. G., Cherayil, J. D., Hampel, A. & Bock, R. M. (1966). *J. Mol. Biol.* **19**, 556.
Söll, D., Ohtsuka, E., Jones, D. S., Lohrmann, R., Hayatsu, H., Nishimura, S. & Khorana, H. G. (1965). *Proc. Nat. Acad. Sci., Wash.* **54**, 1378.